A Decidable Logic for Describing Linked Data Structures

Michael Benedikt* Thomas Reps† Mooly Sagiv‡

Abstract

This paper aims to provide a better formalism for describing properties of linked data structures (e.g., lists, trees, graphs), as well as the intermediate states that arise when such structures are destructively updated. The paper defines a new logic that is suitable for these purposes (called L_r, for "logic of reachability expressions"). We show that L_r is decidable, and explain how L_r relates to two previously defined structure-description formalisms ("path matrices" and "static shape graphs") by showing how an arbitrary shape descriptor from each of these formalisms can be translated into an L_r formula.

Categories and Subject Descriptors: D.3.3 [**Programming Languages**]: Language Constructs and Features—*data types and structures*; *dynamic storage management*; E.1 [**Data**]: Data Structures—*graphs*; *lists*; *trees*; E.2 [**Data**]: Data Storage Representations—*composite structures*; *linked representations*; F.3.1 [**Logics and Meanings of Programs**]: Specifying and Verifying and Reasoning about Programs—*assertions*; *invariants*; *pre- and post-conditions*; *specification languages*; F.3.3 [**Logics and Meanings of Programs**]: Studies of Program Constructs—*type structure*; F.4.1 [**Mathematical Logic and Formal Languages**]: Formal Languages—*decision problems*

General Terms: Algorithms, Languages, Theory

Additional Key Words: Alias information, decidable logic, destructive updating, path matrix, pointer aliasing, program invariant, shape descriptor, static shape graph

1 Introduction

This paper aims to provide a better formalism for describing properties of linked data structures (e.g., lists, trees, graphs). In past work with the same motivation, a variety of different formalisms have been developed — including "static shape graphs" [JM81, JM82, LH88, HPR89, CWZ90, Str92, AW93, PCK93, Wan94, SRW98b, SRW98a], "path matrices" [Hen90, HN90], "graph types" [KS93], and the ADDS annotation formalism [HHN92] — and several previously known formalisms have been exploited — including graph grammars [FM97] and monadic second-order logic [JJNS97]. For lack of a better term, we will use the phrase *structure-description formalisms* to refer to such formalisms in a generic sense.

In this paper, we define a new logic (called L_r, for "logic of reachability expressions"), and show that L_r is suitable for describing properties of linked data structures. We show that L_r is decidable. We also show in detail how L_r relates to two of the previously defined structure-description formalisms: In Section 3, we show how a generalization of Hendren's path-matrix descriptors [Hen90, HN90] can be represented by L_r formulae; in Section 4, we show how the variant of static shape graphs defined in [SRW98b] can be represented by L_r formulae. In this way, L_r provides insight into the expressive power of path matrices and static shape graphs.

The benefits of our work include the following:

*Address: Bell Laboratories; Lucent Technologies; 1000 E. Warrenville Road; P.O. Box 3013; Naperville, IL 60566-7013; USA. E-mail: benedikt@research.bell-labs.com

†Address: Computer Sciences Department; University of Wisconsin; 1210 West Dayton Street; Madison, WI 53706; USA. E-mail: reps@cs.wisc.edu.

‡Address: Dept. of Computer Science; School of Mathematical Sciences; Tel-Aviv University; Tel-Aviv 69978; Israel. E-mail: sagiv@math.tau.ac.il.

- The logic L_r can be used as an annotation language to express loop invariants and pre- and post-conditions of statements and procedures. Annotations are important not only as a means of documenting programs, but also as the basis for analyzing and reasoning about programs in a modular fashion. Our work has two advantages:
  - The logic L_r is quite expressive (e.g., strictly more expressive than the formalism used by Hendren et al. [HHN92]). The added expressibility is important for describing the intermediate states that arise when linked data structures are destructively updated.
  - The logic L_r is *decidable*, which means that there is an algorithm that determines, for every formula in the logic, if the formula is satisfiable. In other words, it is possible to determine if there is any store at all that satisfies a given formula. In principle, this ability can be used to provide some sanity checks on the formulae that a user employs — e.g., a warning can be issued if the user employs a formula that is unsatisfiable.
- Our work makes contributions on the question of *extracting information from the results of program analysis*. Although the subject of the paper is not primarily *algorithms for analyzing* programs that manipulate linked data structures, the decidability of L_r — together with the constructions given in Sections 3 and 4 for encoding other structure-description formalisms in L_r — has interesting consequences for extracting information from the results of program analyses: L_r provides a way to *amplify* the results obtained from known pointer-analysis, alias-analysis, and shape-analysis algorithms in the following ways:
  - For a structure-description formalism in which each structure descriptor corresponds to an L_r formula, as is the case for path matrices (Section 3) and static shape graphs (Section 4), it is possible to determine if there is any store at all that corresponds to a given structure descriptor. This lets us determine whether a given structure descriptor contains any useful information.
  - Pointer-analysis, alias-analysis, and shape-analysis algorithms necessarily compute structure descriptors that over-approximate the pointer/alias/shape relationships that actually arise. This kind of loss of precision is intrinsic to static-analysis; however, many of the techniques that have been proposed in the literature have the feature that *additional imprecision* crops up when information is extracted from the structure descriptor for a particular program point. For instance, with the three-valued logic used for shape analysis in [SRW98a], a formula that queries for a specific piece of information sometimes evaluates to "unknown", even when, in all of the stores that the static shape graph represents, the formula evaluates to a definite true or false value.

    For a structure-description formalism in which each structure descriptor corresponds to an L_r formula, decidability gives us a mechanism for *reading out information obtained by existing algorithms, without any additional loss of precision*: If φ is the formula that represents the shape descriptor and ψ is the formula that represents the query, we are interested in whether $\varphi \implies \psi$ always holds (or, equivalently, whether $\neg(\varphi \implies \psi)$ is unsatisfiable). Thus, in principle, the machinery developed in this paper allows us to take the structure descriptors computed by existing techniques, and extract information from them that is more precise than that envisioned by the inventors of these formalisms.
- For many of the structure-description formalisms used in the literature, very little is known about basic decision problems associated with them. Mapping a structure-description formalism F into L_r can provide a way to analyze many basic decision problems of F.

  For instance, a decision problem of special interest for structure-description formalisms that are used in abstract interpretation is the inclusion problem (i.e., whether the set of stores that structure descriptor D_1 represents is a subset of the set of stores that D_2 represents). When the inclusion problem is decidable, it is possible to check (i) whether one structure descriptor subsumes another (and hence the second need not be retained), and (ii) whether a simpler structure descriptor is a conservative replacement of a larger one, which is useful in widening. Thus, the inclusion problem is important for reducing both the time and space used during abstract interpretation.

  For a structure-description formalism in which each structure descriptor corresponds to an L_r formula, the inclusion of structure descriptor D_1 (represented by formula φ_1) in D_2 (represented by φ_2) is a

```
/* list.h */
typedef struct elem_list_struct {
                int val;
                struct elem_list_struct *cdr; } Elements;
#include "list.h"
Elements * elem_reverse(Elements *x /* acyclic_list(x) */)
{ Elements *z, *y;
  y = NULL ;
  while (x != NULL) {
        /* acyclic_list(x) */
        /* acyclic_list(y) */
        /* disjoint_lists(x,y) */
        z = y;
        y = x ;
        x = x →cdr;
        y →cdr = z; }
  /* acyclic_list(y) */
  return y ;
}
```

Figure 1: A program that uses destructive-updating operations to reverse a list.

matter of testing whether $\varphi_1 \implies \varphi_2$ always holds (or, equivalently, whether $\neg(\varphi_1 \implies \varphi_2)$ is unsatisfiable).

To date, our concern has been with developing the tools for describing properties of linked data structures and obtaining a logic that is decidable. We have developed a decision procedure for L_r, although this procedure does not yield a practical algorithm. We have not yet investigated the complexity of the decision problem for L_r, nor looked for heuristic methods with acceptable performance in practice, but we plan to do so in future work.

Two programs will be used as examples in the paper: Figure 1 shows a program (in a Lisp-like imperative language) that uses destructive-updating operations to reverse a list; Figure 2 shows a program that searches the list pointed to by variable `c` and deletes the first element whose `val` field equals `delval`.

The remainder of the paper is organized into six sections: Section 2 presents the logic we use for describing properties of linked data structures. Section 3 shows how a generalization of Hendren's path-matrix descriptors [Hen90, HN90] can be represented by L_r formulae. Section 4 shows how a variant of static shape graphs can be represented by L_r formulae. Section 5 discusses the issue of using L_r formulae to extract information from the results of program analyses. Section 6 gives a sketch of the proof that L_r is decidable. Section 7 discusses related work.

2 A Language for Stating Properties of Linked Data Structures

Definition 2.1 *Let PVar be the (finite) set of pointer variables in a given program. Let Sel be the set of* **pointer selectors** *(i.e., pointer-valued fields of structures) used in the program. We define the* **alphabet** *Σ to be the following finite set of symbols:*

$$\Sigma = Sel \cup \{\texttt{pvar?} \mid \texttt{pvar} \in PVar\} \cup \{\neg\texttt{pvar?} \mid \texttt{pvar} \in PVar\}$$

```
/* list.h */
typedef struct elem_list_struct {
                int val;
                struct elem_list_struct *cdr; } Elements;
#include "list.h"
bool elem_delete(int delval, Elements *c)
{ /* acyclic_list(c) */
  Elements *elem,*prev;
  for (prev = NULL, elem = c;
       elem != NULL;
       prev=elem, elem = elem →cdr) {
      if (elem →val == delval) {
         if (prev == NULL)
            c = elem →cdr;
         else
               prev →cdr = elem →cdr;
         free(elem);
         return TRUE; } }
  return FALSE;
/* acyclic_list(c) */
}
```

Figure 2: A program that searches the list pointed to by variable `c` (using a "trailing pointer" `prev`) and deletes the first element whose `val` field equals `delval`.

with the intended meaning that $\mathtt{pvar}?$ *(*$\neg\mathtt{pvar}?$*) denotes cells pointed to by (not pointed to by) a pointer variable* `pvar`. *A* **formula** *in the logic* L_r *is defined as follows:*

Φ	$::=$	$\mathtt{pe}_1 = \mathtt{pe}_2$	*equality of pointer expressions*
	$\mid$	$\mathtt{pe}_1\langle R\rangle\mathtt{pe}_2$	*reachability constraint*
	$\mid$	$hs(\mathtt{pe}\langle R\rangle)$	*heap-sharing constraint*
	$\mid$	$al(\mathtt{pe}\langle R\rangle)$	*allocation constraint*
	$\mid$	$\neg\Phi$	*negation*
	$\mid$	$\Phi_1 \wedge \Phi_2$	*conjunction*
	$\mid$	$\Phi_1 \vee \Phi_2$	*disjunction*
	$\mid$	$\Phi_1 \implies \Phi_2$	*implication*
$\mathtt{pe}$	$::=$	$\mathtt{pvar}$	*pointer variable*
	$\mid$	$\mathtt{pe.sel}$	$\mathtt{sel} \in Sel$
R	$::=$	ϵ	*empty path*
	$\mid$	$\emptyset$	*empty language*
	$\mid$	σ	$\sigma \in \Sigma$
	$\mid$	$R_1.R_2$	*concatenation*
	$\mid$	$R_1 \mid R_2$	*union*
	$\mid$	R^*	*Kleene star*

We call R *terms* **routing expressions**, *and refer to occurrences of* $\mathtt{pvar}?$ *and* $\neg\mathtt{pvar}?$ *in routing expressions as* **pointer-variable interrogations**.

We also use several shorthand notations: $hs(\mathtt{p})$ *and* $al(\mathtt{p})$ *are shorthands for* $hs(\mathtt{p}\langle\epsilon\rangle)$ $al(\mathtt{p}\langle\epsilon\rangle)$, *respectively. Similarly,* $hs(\mathtt{p.sel})$ *and* $al(\mathtt{p.sel})$ *are shorthands for* $hs(\mathtt{p}\langle\mathtt{sel}\rangle)$ $al(\mathtt{p}\langle\mathtt{sel}\rangle)$, *respectively.* $\mathtt{pe}_1 \neq \mathtt{pe}_2$ *is a shorthand for* $\neg(\mathtt{pe}_1 = \mathtt{pe}_2)$. $\Phi_1 \iff \Phi_2$ *is a shorthand for* $(\Phi_1 \implies \Phi_2) \wedge (\Phi_2 \implies \Phi_1)$. □

Example 2.2 For a pointer variable x, the formula

$$acyclic\_list(\mathtt{x}) \stackrel{\text{def}}{=} \neg\mathtt{x}\langle\mathtt{cdr}^+\rangle\mathtt{x} \wedge \neg hs(\mathtt{x}\langle\mathtt{cdr}^*\rangle)$$

states that x points to an unshared acyclic list. The term $\neg\mathtt{x}\langle\mathtt{cdr}^+\rangle\mathtt{x}$ signifies that a traversal that starts with the cell pointed to by x and follows one or more cdr fields never returns to the cell pointed to by x. The term $\neg hs(\mathtt{x}\langle\mathtt{cdr}^*\rangle)$ signifies that a traversal that starts with the cell pointed to by x and follows zero or more cdr fields never leads to a cell that is "heap shared". (A cell c is "heap shared" if two or more cells have fields that point to c, or if there is a cell c' such that c'.car and c'.cdr both point to c [JM82, CWZ90, SRW98b, SRW98a].)

Thus, a loop invariant for program elem_reverse can be written as follows:

$$\begin{array}{l} ((al(\mathtt{y.cdr}) \vee al(\mathtt{z})) \implies \mathtt{y.cdr} = \mathtt{z}) \\ \wedge\ acyclic\_list(\mathtt{x}) \wedge acyclic\_list(\mathtt{y}) \\ \wedge\ \neg\mathtt{x}\langle\mathtt{cdr}^*\rangle\mathtt{y} \wedge \neg\mathtt{x}\langle\mathtt{cdr}^*\rangle\mathtt{z} \wedge \neg\mathtt{y}\langle\mathtt{cdr}^*\rangle\mathtt{x} \wedge \neg\mathtt{z}\langle\mathtt{cdr}^*\rangle\mathtt{x} \end{array} \tag{1}$$

The first line of (1) states that y.cdr and z refer to the same list element when either one is allocated. The subformulae on the last line of (1) states that the x-list is disjoint from both the y-list and the z-list. □

Example 2.3 A loop invariant for program elem_delete can be written as follows:

$$\begin{array}{l} acyclic\_list(\mathtt{c}) \\ \wedge\ \mathtt{c}\langle\mathtt{cdr}^*\rangle\mathtt{elem} \\ \wedge\ al(\mathtt{prev}) \implies (\mathtt{c}\langle\mathtt{cdr}^*\rangle\mathtt{prev} \wedge \mathtt{prev.cdr} = \mathtt{elem}) \\ \wedge\ \neg al(\mathtt{prev}) \iff \mathtt{elem} = \mathtt{c} \end{array} \tag{2}$$

The subformula $\mathtt{c}\langle\mathtt{cdr}^*\rangle\mathtt{elem}$ in the second line of (2) states that elem points somewhere in the list pointed to by c. The subformula on the fourth line of (2) implies that when prev is allocated (i.e., not NULL), elem and c must point to different locations. From this, we can conclude that the location released by the statement free(elem) cannot be pointed to by c. □

The use of pointer-variable interrogations in routing expressions will be illustrated in Examples 3.3 and 3.6.

We now define the semantics of L_r formulae:

Definition 2.4 *A store S can be represented by a tuple $\langle Loc^S, env^S, \iota^S\rangle$, where Loc^S is a set of locations, and env^S and ι^S are functions*

$$\begin{array}{rcl} env^S & : & PVar^S \to (Loc^S \cup \{\perp\}) \\ \iota^S & : & Sel^S \to (Loc^S \cup \{\perp\}) \to (Loc^S \cup \{\perp\}), \end{array}$$

where ι^S is strict in $\perp$.

The meaning of a pointer expression pe *in a given store S, denoted by $[\![\mathtt{pe}]\!]^S$ (where $[\![\mathtt{pe}]\!]^S \in (Loc^S \cup \{\perp\})$), is defined inductively, as follows:*

$$\begin{array}{rcl} [\![\mathtt{pvar}]\!]^S & = & env^S(\mathtt{pvar}) \\ [\![\mathtt{pe.sel}]\!]^S & = & \iota^S(\mathtt{sel})([\![\mathtt{pe}]\!]^S) \end{array}$$

The language $L(R)$ of a routing expression R is defined as is usual for regular expressions. However, because a word in $L(R)$ can contain occurrences of pointer-variable interrogations of the form pvar? *and* ¬pvar?*, the meaning of a word is slightly nonstandard: The meaning of a word in a given store S, denoted by $[\![w]\!]^S$ (where $[\![w]\!]^S : (Loc^S \cup \{\perp\}) \to (Loc^S \cup \{\perp\})$), is defined inductively, as follows:*

$$\begin{array}{rcl} [\![\epsilon]\!]^S(l) & = & l \\ [\![w.\mathtt{sel}]\!]^S(l) & = & \iota^S(\mathtt{sel})([\![w]\!]^S(l)) \\ [\![w.\mathtt{pvar?}]\!]^S(l) & = & \begin{cases} [\![w]\!]^S(l) & \text{if } env(\mathtt{pvar}) = [\![w]\!]^S(l) \\ \perp & \text{otherwise} \end{cases} \\ [\![\neg\mathtt{sel?}]\!]^S(l) & = & \begin{cases} [\![w]\!]^S(l) & \text{if } env(\mathtt{pvar}) \neq [\![w]\!]^S(l) \\ \perp & \text{otherwise} \end{cases} \end{array}$$

The meaning of formula in a given store S is defined inductively, as follows:

$$
\begin{array}{rcl}
[\![\mathtt{pe}_1 = \mathtt{pe}_2]\!]^S & = & ([\![\mathtt{pe}_1]\!]^S = [\![\mathtt{pe}_2]\!]^S) \\
[\![\mathtt{pe}_1 \langle R \rangle \mathtt{pe}_2]\!]^S & = & \text{there exists } w \in L(R) \text{ such that } [\![w]\!]^S([\![\mathtt{pe}_1]\!]^S) = [\![\mathtt{pe}_2]\!]^S \text{ and } [\![\mathtt{pe}_2]\!]^S \in \mathit{Loc} \\
[\![hs(\mathtt{pe}\langle R \rangle)]\!]^S & = & \text{there exists } w \in L(R) \text{ such that } [\![w]\!]^S([\![\mathtt{pe}]\!]^S) = l \text{ and } l \in \mathit{Loc} \text{ and} \\
 & & \text{there exist } l_1, l_2 \in \mathit{Loc}, \mathtt{sel}_1, \mathtt{sel}_2 \in \mathit{Sel} \text{ such that } \iota^S(\mathtt{sel}_1)(l_1) = l \text{ and } \iota^S(\mathtt{sel}_2)(l_2) = l \text{ and} \\
 & & \text{either (i) } l_1 \neq l_2 \text{ or (ii) } \mathtt{sel}_1 \neq \mathtt{sel}_2 \\
[\![al(\mathtt{pe}\langle R \rangle)]\!]^S & = & \text{there exists } w \in L(R) \text{ such that } [\![w]\!]^S([\![\mathtt{pe}]\!]^S) \in \mathit{Loc} \\
[\![\neg\Phi]\!]^S & = & [\![\Phi]\!]^S \text{ is false} \\
[\![\Phi_1 \wedge \Phi_2]\!]^S & = & [\![\Phi_1]\!]^S \text{ is true and } [\![\Phi_2]\!]^S \text{ is true} \\
[\![\Phi_1 \vee \Phi_2]\!]^S & = & [\![\Phi_1]\!]^S \text{ is true or } [\![\Phi_2]\!]^S \text{ is true} \\
[\![\Phi_1 \implies \Phi_2]\!]^S & = & [\![\neg\Phi_1]\!]^S \text{ is true or } [\![\Phi_2]\!]^S \text{ is true}
\end{array}
$$

□

3 Representing Path Matrices via Formulae

In this section, we study the relationship between the logic L_r and a variant of the *path-matrix* structure-description formalism [Hen90, HN90]. A path matrix records information about the (possibly empty) set of paths that exist between pairs of pointer variables in a program. The version of path matrices described below is a generalization of the original version described in [Hen90, HN90]. We show that every path matrix (of the extended version of the formalism) can be represented by a formula in logic L_r.

Definition 3.1 *A path matrix pm contains an entry $pm[\mathtt{x},\mathtt{y}]$ for every pair of pointer-valued program variables, $\mathtt{x}$ and $\mathtt{y}$. An entry $pm[\mathtt{x},\mathtt{y}]$ describes the set of paths from the cell pointed to by $\mathtt{x}$ to the cell pointed to by $\mathtt{y}$. An entry $pm[\mathtt{x},\mathtt{y}]$ has a value of the form $\langle R, Q\rangle$, where R is a regular expression over Σ, and Q is either "P" (standing for "possible path") or "D" (standing for "definite" path).* □

The notions of "possible paths" and "definite paths" are somewhat subtle (and the names "possible paths" and "definite paths", which we have adopted from [Hen90, HN90], are somewhat misleading). In the discussion below, let S be a store that path matrix pm represents, and let $\mathit{Paths}_S(\mathtt{x},\mathtt{y})$ denote the set of paths from the cell pointed to by program variable $\mathtt{x}$ to the cell pointed to by $\mathtt{y}$.

- An entry $pm[\mathtt{x},\mathtt{y}]$ that has the value $\langle R_D, D\rangle$ means that there is a path p in S from the cell pointed to by program variable $\mathtt{x}$ to the cell pointed to by $\mathtt{y}$, such that $p \in L(R_D)$. In other words,

  $$\mathit{Paths}_S(\mathtt{x},\mathtt{y}) \cap L(R_D) \neq \emptyset. \tag{3}$$

  Note that only *one* of the paths in $L(R_D)$ need be a path in $\mathit{Paths}_S(\mathtt{x},\mathtt{y})$ for $pm[\mathtt{x},\mathtt{y}] = \langle R_D, D\rangle$ to be satisfied.

- An entry $pm[\mathtt{x},\mathtt{y}]$ that has the value $\langle R_P, P\rangle$ means that $L(R_P)$ is an over-approximation to the set of paths from the cell pointed to by $\mathtt{x}$ to the cell pointed to by $\mathtt{y}$. In other words,

  $$\mathit{Paths}_S(\mathtt{x},\mathtt{y}) \subseteq L(R_P). \tag{4}$$

  An alternative way to think about this is as follows: What we really mean by "R_P represents possible paths in store S" is that $\overline{L(R_P)} = \Sigma^* - L(R_P)$ is a set of *impossible* paths of S: That is, an entry $pm[\mathtt{x},\mathtt{y}]$ that has the value $\langle R_P, P\rangle$ means that none of the paths from the cell pointed to by $\mathtt{x}$ to the cell pointed to by $\mathtt{y}$ are in $\overline{L(R_P)}$. Thus, we have

  $$\mathit{Paths}_S(\mathtt{x},\mathtt{y}) \cap \overline{L(R_P)} = \emptyset. \tag{5}$$

These two ways of looking at things are equivalent, as shown by the following derivation:

$$\begin{array}{rcl} \mathit{Paths}_S(\mathtt{x},\mathtt{y}) \cap \overline{L(R_P)} & = & \emptyset \\ \mathit{Paths}_S(\mathtt{x},\mathtt{y}) - L(R_P) & = & \emptyset \\ \mathit{Paths}_S(\mathtt{x},\mathtt{y}) & \subseteq & L(R_P). \end{array}$$

It is instructive to consider some simple examples of possible path-matrix entries:

- An entry $pm[\mathtt{x},\mathtt{y}]$ that has the value $\langle \emptyset, P\rangle$ represents the fact that there is no path in S from the cell pointed to by program variable `x` to the cell pointed to by `y`.
- An entry $pm[\mathtt{x},\mathtt{y}]$ that has the value $\langle \epsilon, D\rangle$ represents the fact that `x` and `y` are must-aliases, i.e., `x` and `y` must point to the same cell in all of the stores that the path matrix represents.
- In contrast, an entry $pm[\mathtt{x},\mathtt{y}]$ with the value $\langle \epsilon, P\rangle$ represents the fact that `x` and `y` are may-aliases, i.e., `x` and `y` might point to the same cell in some of the stores that the path matrix represents, but it is also possible that in other stores that the path matrix represents, there is no path at all from the cell pointed to by `x` to the cell pointed to by `y`.
- More generally, a value $\langle R, P\rangle$ for entry $pm[\mathtt{x},\mathtt{y}]$, where $\epsilon \in L(R)$ means that `x` and `y` are may aliases. The language $L(R) - \{\epsilon\}$ represents other possible paths from the cell pointed to by `x` to the cell pointed to by `y`, but it is also possible that in some of the stores that the path matrix represents, there is no path at all from the cell pointed to by `x` to the cell pointed to by `y`.

Note that a path matrix represents a smaller set of stores if the language for a "D" entry is made smaller, and also if the language for a "P" entry is made smaller (see (3) and (4)).

Example 3.2 The following path matrix expresses a loop-invariant for the loop of `elem_reverse`:

pm	`x`	`y`	`z`
`x`	$\langle \epsilon, D\rangle$	$\langle \emptyset, P\rangle$	$\langle \emptyset, P\rangle$
`y`	$\langle \emptyset, P\rangle$	$\langle \epsilon, D\rangle$	$\langle \mathtt{cdr}, D\rangle$
`z`	$\langle \emptyset, P\rangle$	$\langle \emptyset, P\rangle$	$\langle \epsilon, D\rangle$

(6)

The fact that $pm[\mathtt{y},\mathtt{z}]$ is $\langle \mathtt{cdr}, D\rangle$ signifies that $\mathtt{y} \rightarrow \mathtt{cdr}$ must point to the cell that `z` points to. □

Example 3.3 The following path matrix expresses a loop-invariant for the loop of `elem_delete`:

pm	`prev`	`elem`	`c`
`prev`	$\langle \epsilon, D\rangle$	$\langle \mathtt{cdr}, P\rangle$	$\langle \emptyset, P\rangle$
`elem`	$\langle \emptyset, P\rangle$	$\langle \epsilon, D\rangle$	$\langle \emptyset, P\rangle$
`c`	$\langle \mathtt{cdr}^*, P\rangle$	$\langle \epsilon \vert \mathtt{cdr}^*.\mathtt{prev?}.\mathtt{cdr}, P\rangle$	$\langle \epsilon, D\rangle$

(7)

The fact that $pm[\mathtt{prev},\mathtt{elem}]$ is $\langle \mathtt{cdr}, P\rangle$ signifies that $\mathtt{prev} \rightarrow \mathtt{cdr}$ may point to the cell pointed to by `elem`, but may also point to a cell that `elem` does not point to; in fact, the latter is the case at the beginning of the first loop iteration. Similarly, the fact that $pm[\mathtt{c},\mathtt{prev}]$ is $\langle \mathtt{cdr}^*, P\rangle$ signifies that `prev` may be reachable from `c`. The fact that $pm[\mathtt{c},\mathtt{elem}]$ entry is $\langle \epsilon \vert \mathtt{cdr}^*.\mathtt{prev?}.\mathtt{cdr}, P\rangle$ signifies that either `c` and `elem` point to the same cell, or else that as we traverse the list pointed to by `c`, we first reach a cell pointed to by `prev` and then the cell pointed to by `elem`. □

Remark. The routing expressions that we allow in path matrices are more general than the ones allowed in [Hen90, HN90] in the following way:

- We allow arbitrary alternations and not just `car|cdr`.
- We follow [JJNS97] in allowing pointer-variable interrogations (e.g., `prev`, $\neg$`prev`) in routing expressions. This comes in handy in cases where several paths depend on each other (cf. the $pm[\mathtt{c},\mathtt{elem}]$ entry in path matrix (7)).

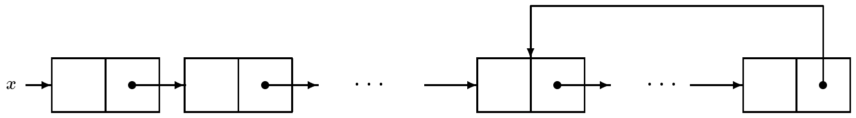

Figure 3: A store with a shared node.

Of course these limitations of [Hen90, HN90] are motivated by the need to be able to compute efficiently a safe approximation to the path matrix at every program point. □

Since path matrices are an intuitive notation, we will not spend the space in directly formalizing the meaning of path matrices in terms of sets of stores. Instead, we now define the meaning of a path matrix by a formula in our language that characterizes the set of stores that a path matrix represents.

Definition 3.4 *For a regular expression R, let $\overline{R}$ denote the complement of R, i.e., a regular expression such that $L(\overline{R}) = \overline{L(R)} = \Sigma^* - L(R)$. For a path matrix pm, we define the formula φ_{pm} as follows:*

$$\varphi_{pm} \stackrel{\text{def}}{=} \bigwedge_{\mathtt{x},\mathtt{y} \in PVar, \langle R,D\rangle \in pm[\mathtt{x},\mathtt{y}]} \mathtt{x}\langle R\rangle\mathtt{y} \quad\wedge \bigwedge_{\mathtt{x},\mathtt{y} \in PVar, \langle R,P\rangle \in pm[\mathtt{x},\mathtt{y}]} \neg\mathtt{x}\langle \overline{R}\rangle\mathtt{y} \tag{8}$$

□

This definition is justified by the discussion that follows Definition 3.1.

Example 3.5 Path matrix (6), which expresses a loop-invariant for the loop of `elem_reverse` (see Example 3.2), corresponds to the following formula:

$$\begin{array}{rrrr} & \mathtt{x}\langle\epsilon\rangle\mathtt{x} & \wedge\ \neg\mathtt{x}\langle\Sigma^*\rangle\mathtt{y} & \wedge\ \neg\mathtt{x}\langle\Sigma^*\rangle\mathtt{z} \\ \wedge & \neg\mathtt{y}\langle\Sigma^*\rangle\mathtt{x} & \wedge\quad \mathtt{y}\langle\epsilon\rangle\mathtt{y} & \wedge\ \mathtt{y}\langle\mathtt{cdr}\rangle\mathtt{z} \\ \wedge & \neg\mathtt{z}\langle\Sigma^*\rangle\mathtt{x} & \wedge\ \neg\mathtt{z}\langle\Sigma^*\rangle\mathtt{y} & \wedge\quad \mathtt{z}\langle\epsilon\rangle\mathtt{z} \end{array} \tag{9}$$

Formula (9) is less informative than the loop-invariant given as Formula (1) of Example 2.2. For example, with Formula (9) it is not known that `x` points to a list, because cyclic stores of the form shown in Figure 3 also satisfy (9). □

Example 3.6 Path matrix (10), which expresses a loop-invariant for the loop of `elem_delete` (see Example 3.3), corresponds to the following formula:

$$\begin{array}{rrrr} & \mathtt{prev}\langle\epsilon\rangle\mathtt{prev} & \wedge\quad \neg\mathtt{prev}\langle\overline{\mathtt{cdr}}\rangle\mathtt{elem} & \wedge\ \neg\mathtt{prev}\langle\Sigma^*\rangle\mathtt{c} \\ \wedge & \neg\mathtt{elem}\langle\Sigma^*\rangle\mathtt{prev} & \wedge\quad \mathtt{elem}\langle\epsilon\rangle\mathtt{elem} & \wedge\quad \mathtt{elem}\langle\Sigma^*\rangle\mathtt{c} \\ \wedge & \neg\mathtt{c}\langle\overline{\mathtt{cdr}^*}\rangle\mathtt{prev} & \wedge\ \neg\mathtt{c}\langle\overline{\epsilon|\mathtt{cdr}^*.\mathtt{prev?}.\mathtt{cdr}}\rangle\mathtt{elem} & \wedge\quad \mathtt{c}\langle\epsilon\rangle\mathtt{c} \end{array} \tag{10}$$

Formula (10) is less informative than the loop-invariant given as Formula (2) of Example 2.3. In contrast to Formula (2), Formula (10) cannot be used to conclude that the use of `free` in `elem_delete` is correct; i.e., we cannot conclude that the location released by the statement `free(elem)` cannot be pointed to by `c`. □

4 Representing Shape Graphs via Formulae

In this section, we study a structure-description formalism called *static shape graphs*, which, in addition to reachability information, allow certain "topological" properties of stores to be represented. There are many ways to define static shape graphs. For simplicity, we only consider the variant of static shape graphs defined in [SRW98b].

The remainder of this section is organized as follows. In Section 4.1, we give a formal definition of static shape graphs. Then, in Section 4.2, we construct a formula in L_r that exactly characterizes the set of stores represented by a static shape graph.

4.1 Static Shape Graphs

Below, we formally define static shape graphs. Unlike the stores defined in Section 2, static shape graphs are of an *a priori* bounded size, i.e., the number of shape nodes depends only of the size of the program being analyzed. This is needed by shape-analysis algorithms so that an iterative shape-analysis algorithm that computes static shape graphs for each program point will terminate.

Definition 4.1 *A* **static-shape-graph** *(SSG) is a finite directed graph that consists of two kinds of nodes —* **variables** *(i.e., PVar) and* **shape-nodes** *— and two kinds of edges —* **variable-edges** *and* **selector-edges**. *A shape-graph is represented by a quadruple $\langle shapeNodes, E_v, E_s, is \rangle$, where:*

- *shapeNodes is a finite set of shape nodes. Every shape node $n \in ShapeNodes$ has the form $n = n_X$ where $X \subseteq PVar$. Such a node describes the cells that are simultaneously pointed to by all the pointer variables in X.*

  Graphically, we denote shape nodes by circles. The node $n_\emptyset$ is the "summary-node" since it represents all the cells that are not directly pointed by pointer variables, and therefore it is represented by a dotted circle.

- *E_v is the graph's set of variable-edges, each of which is denoted by a pair of the form $[\mathtt{x}, n_X]$, where $\mathtt{x} \in PVar$ and $n_X \in shapeNodes$. We assume that for every $\mathtt{x} \in PVar$, at most one variable-edge $[\mathtt{x}, n_X] \in E_v$ exists and $\mathtt{x} \in X$.*

  Graphically, we denote variable-edges by solid edges since they must exist.

- *E_s is the graph's set of selector-edges, each of which is denoted by a triple of the form $\langle n_X, sel, n_Y \rangle$, where $n_X, n_Y \in shapeNodes$ and $sel \in \{car, cdr\}$. We assume that for every $\mathtt{x} \in PVar$, $sel \in \{car, cdr\}$, and shape node n_X such that $[\mathtt{x}, n_X] \in E_v$, at most one selector-edge, $\langle n_X, sel, n_Y \rangle \in E_s$ exists. In contrast, there may be many selector-edges $\langle n_\emptyset, sel, n_Y \rangle \in E_s$ corresponding to different selector-edges emanating from cells represented by $n_\emptyset$.*

  Graphically, we denote selector-edges by dotted edges since they may or may not exist.

- *is is a function of type $shapeNodes \rightarrow \{false, true\}$. It serves as a constraint to restrict the set of stores represented by a shape graph. When $n_\emptyset$ has more than one incoming selector edge and yet $is(n_\emptyset) = false$, we know that, for any memory cell c represented by $n_\emptyset$, at most one of the concrete representatives of these selector-edges can be an incoming edge of c.*

  Graphically, we denote the fact that n_X is a shared node by putting "$is(n_X)$" inside the circle.

□

Example 4.2 The SSGs that arise at the loop header in `elem_reverse` are shown in Figure 4. In all of the SSGs in Figure 4, for all shape nodes n_X, $is(n_X) = false$. The first line in Figure 4 shows the three SSGs that arise before the first iteration of the analysis. These SSGs describe stores in which `x` points to an empty list, `x` points to a list of length one, and `x` points to a list of length two or more, respectively. □

Example 4.3 The SSG that represents the store shown in Figure 3 is shown in Figure 5. □

4.2 From a Static Shape Graph into an L_rFormula

We are now ready to show how to construct the formula that captures the meaning of a static shape graph.

Definition 4.4 *Let $SG = \langle shapeNodes, E_v, E_s, is \rangle$ be an SSG. We define the graph $\widehat{SG}$ to be a directed graph, $\widehat{SG} = (N, A, l)$, with edges labeled by letters in Σ, where:*

- *N contains two nodes $u.in$ and $u.out$ for every shape node u in shapeNodes.*
- *$A \subseteq N \times N$ contains the following two types of labeled edges:*

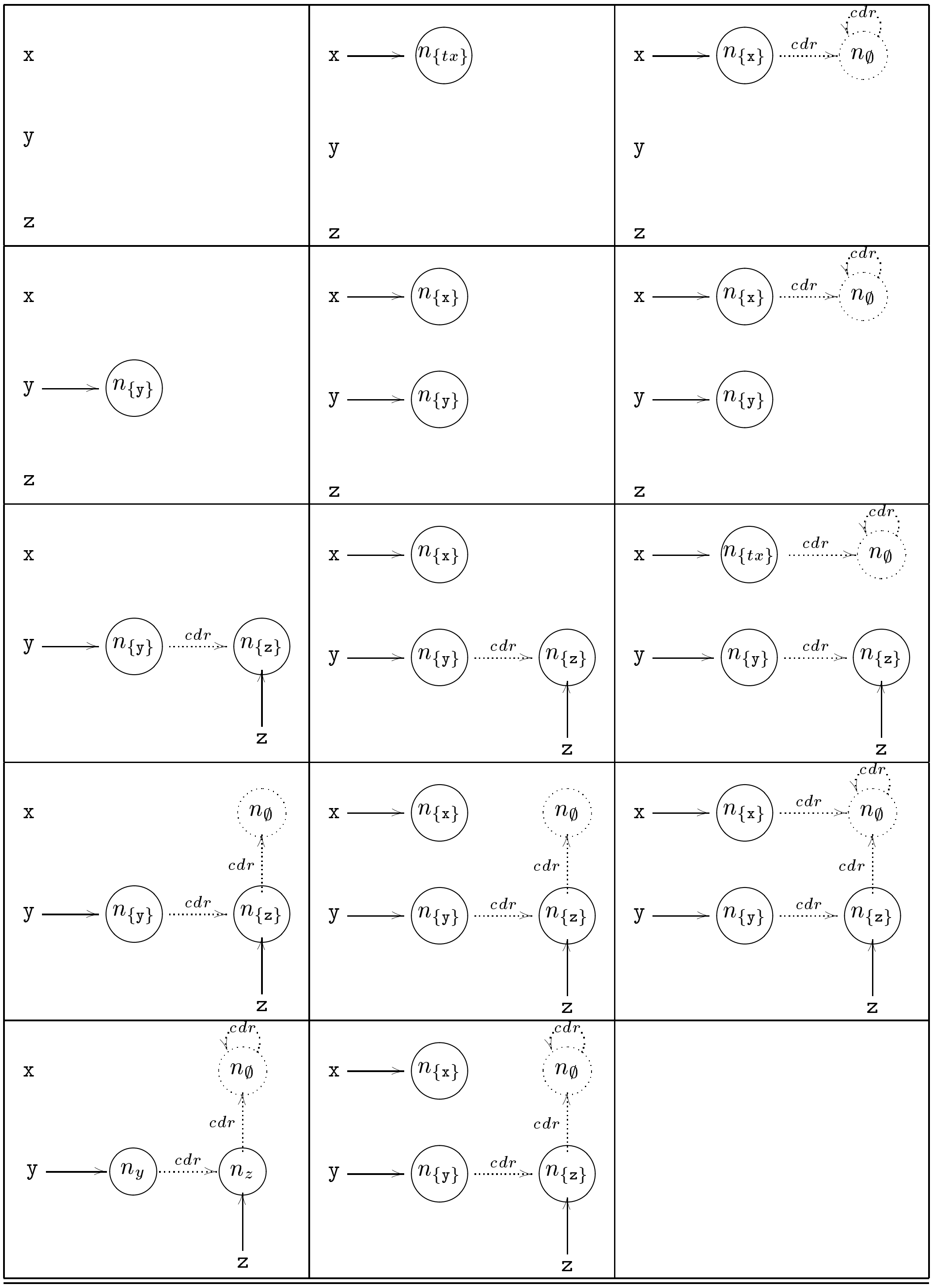

Figure 4: The SSGs that arise at the loop header in `elem_reverse`. In all of the SSGs, for all shape nodes n_X, $is(n_X) = \mathit{false}$.

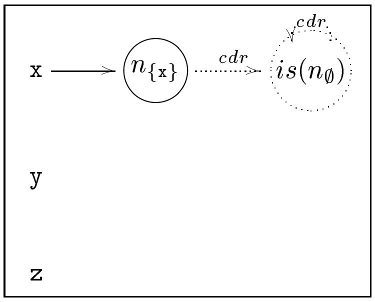

Figure 5: The SSG that corresponds to the store shown in Figure 3.

- *For every shape node* n_X *such that such that* $[p_1, n_X], [p_2, n_X], \cdots, [p_n, n_X] \in E_v$ *and* $[p_{n+1}, n_X], [p_{n+2}, n_X], \cdots, [p_{n+k}, n_X] \notin E_v$, *there is an edge* $\langle n_X.in, n_Y.out\rangle$, *labeled* $(p_1?.p_2?.\ldots.p_n?.\neg p_{n+1}?.\neg p_{n_2}.\neg p_{n+k})$.
- *If there is a selector-edge* $\langle n_X, sel, n_Y\rangle \in E_s$, *there is an edge* $a = (n_X.out, n_Y.in)$ *from* $n_X.out$ *into* $N_Y.in$, *labeled* sel.

- $l\colon A \to \Sigma$ *maps edges into their labels.*

For any two nodes $n_X, n_Y \in shapeNodes$, *let* $r_{n_X.in \to n_Y.out}$ *be the regular path expression over* Σ *that describes paths in* $\widehat{SG}$ *from* $n_X.in$ *into* $n_Y.out$ *(which can be computed by well-known methods, e.g., [Tar81b, Tar81a]). Also, let* $\overline{r_{n_X.in \to n_Y.out}}$ *be the regular path expression over* Σ *that describes the non-existing paths from* $n_X.in$ *into* $n_Y.out$, *i.e.,*

$$L(\overline{r_{n_X.in \to n_Y.out}}) = \Sigma^* - L(r_{n_Y.in \to n_Y.out}).$$

Let us define the following formulae to characterize the different aspects of SG

$$\Phi^+_{val} = \bigwedge_{\mathtt{x} \in PVar, [\mathtt{x}, n_X] \in E_v} al(\mathtt{x}) \tag{11}$$

$$\Phi^-_{val} = \bigwedge_{\mathtt{x} \in PVar, [\mathtt{x}, n_X] \notin E_v} \neg al(\mathtt{x}) \tag{12}$$

$$\Phi^+_{veq} = \bigwedge_{\mathtt{x},\mathtt{y} \in PVar, [\mathtt{x}, n_X], [\mathtt{y}, n_X] \in E_v} \mathtt{x} = \mathtt{y} \tag{13}$$

$$\Phi^-_{veq} = \bigwedge_{\mathtt{x},\mathtt{y} \in PVar, [\mathtt{x}, n_X] \in E_v, [\mathtt{y}, n_X] \notin E_v} \mathtt{x} \neq \mathtt{y} \tag{14}$$

$$\Phi^-_{pal} = \bigwedge_{\mathtt{x} \in PVar, [\mathtt{x}, n_X] \in E_v} \neg al(\mathtt{x}\langle\overline{\Sigma_{n_Y \in shapeNodes} r_{n_X.in \to n_Y.out}}\rangle) \tag{15}$$

$$\Phi^-_{r} = \bigwedge_{[\mathtt{x}, n_X], [\mathtt{y}, n_Y] \in E_v} \neg \mathtt{x}\langle\overline{r_{n_X.in \to n_Y.out}}\rangle\mathtt{y} \tag{16}$$

$$\Phi^-_{hs} = \bigwedge_{\mathtt{x} \in PVar, [\mathtt{x}, n_X] \in E_v} \neg hs(\mathtt{x}\langle\overline{\Sigma_{n_Y \in shapeNodes, is(v)=1} r_{n_X.in \to n_Y.out}}\rangle) \tag{17}$$

Finally, the formula $\Phi[SG]$ *is the conjunction of these formulae.* □

Lemma 4.5 *For every store* S *and SSG* SG, S *is represented by* SG *if and only if* $[\![\Phi[SG]]\!]^S$ *is true.* □

5 Extracting Information from Program Analyses via L_r Formulae

Many interesting properties of linked data structures can be expressed as L_r formulae:

- For example, the formula

$$\Psi \stackrel{\text{def}}{=} (x = x.cdr)$$

expresses the property "x points to a cell that has a self-cycle". This information can be used by an optimizing compiler to determine whether it is profitable to generate a pre-fetch for the next element [LM96].

- It is possible to express in L_r that two pointer-access paths point to different memory cells (i.e., they are not may-aliases), which is important both for optimization and in tools for aiding software understanding.

- The reachability and sharing predicates can also be useful, for example, to improve the performance of garbage-collection algorithms and to parallelize programs.

In principle, L_r provides a uniform basis for using the results of analyses that yield either path matrices or static shape graphs in program optimizers and in tools for aiding software understanding. For instance, Ψ is not satisfiable by any store that is represented by the SSG SG shown in column three of the first line of Figure 4, which means that x does not point to a cell that has a self-cycle in any of the stores that SG represents. This can be determined automatically with our approach by showing that $\Phi[SG] \wedge \Psi$ is not satisfiable. Similarly, by translating a path matrix M (obtained from a path-matrix-based program-analysis algorithm) into the corresponding L_r formula $\Phi[M]$ and checking whether $\Phi[M] \wedge \Psi$ is satisfiable, one can verify automatically whether x could point to a cell that has a self-cycle in any of the stores represented by M.

6 The Decidability of L_r

Theorem 6.1 *L_r is decidable.*
Sketch of Proof: Prior to directly approaching the question of decidability of L_r, one first proves a normalization lemma showing that the routing expressions mentioned in formulae can be rewritten in such a way that they deal only with paths that avoid all nodes pointed by pointer expressions that are mentioned in the formula (i.e. pointer expressions that occur in some constraint or program variables that occur in some pointer-variable interrogation). That is, they assert only reachability of shared nodes or pointer expressions via paths that traverse nodes in the heap. One proves this normalization lemma by breaking down path expressions that may cross mentioned pointer expressions into component path expressions that do not cross mentioned pointer expressions.

The decidability of logic L_r follows from showing that that L_r has the *bounded model property*: that is, there is a computable numerical function f such that any sentence ϕ of L_r that is consistent has a model of size bounded by $f(|\phi|)$. This technique is one of the most common in logical decision procedures [BGG96]. It immediately shows the existence of a crude decision procedure: one enumerates all possible stores of size $f(|\phi|)$ searching for a model. (Note that the approach sketched here is intended only to give a comprehensible demonstration of decidability, not to give a practical decision procedure.) The proof of the bounded model property proceeds by starting with an arbitrary concrete store G satisfying a formula ϕ and showing that G can be diminished to a model of size $|f(|\phi|)$ (for a particular f given in the proof) while preserving all atomic formulae in ϕ.

The normalization theorem above implies that in this shrinking process one only has to preserve properties that deal with paths through the heap (reachability, heap-sharing, etc.) and equalities and inequalities between a fixed set of pointer expressions. This shrinking is then done in three phases: first, the original store G is "pruned" to get a model that is a union of trees: in the process, some information about the sharing of nodes is lost, but extra labels are added to the nodes to maintain this information. These "auxiliary labels" indicate that certain nodes in the tree correspond to nodes associated with a particular pointer expression in the original store, and that certain nodes in the tree were shared in the original store.

We then make use of classical decidability results on reachability expressions on finite trees ([TW68], summarized also in [BGG96]) to shrink each of these trees to smaller trees that satisfy the same properties as the union of trees produced in stage one. The "properties" mentioned here are obtained by taking

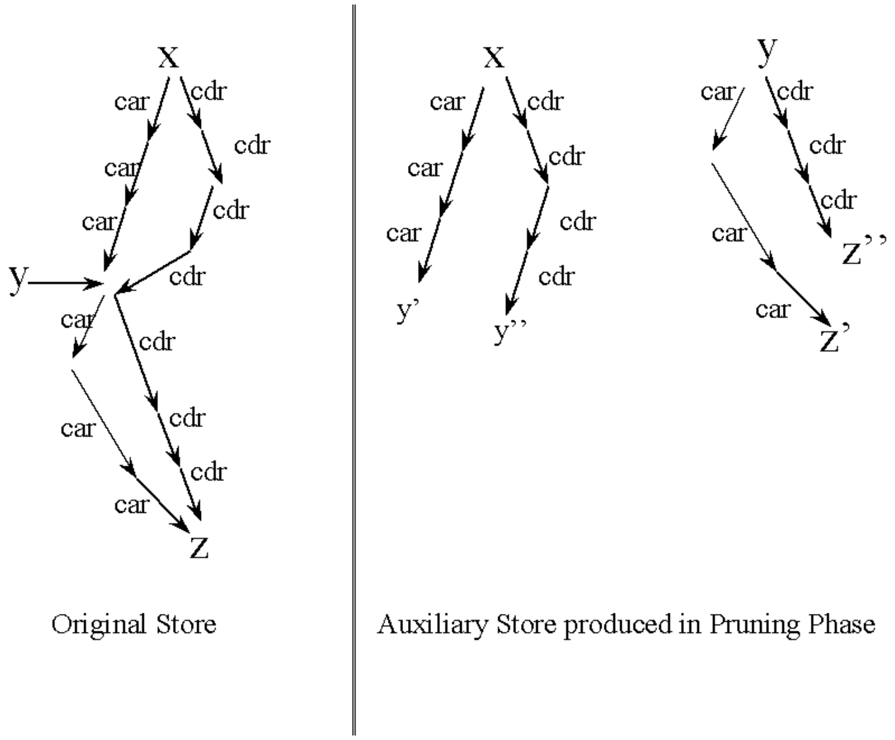

Figure 6: The Pruning Stage of the Proof

the original reachability, heap-sharing, and allocation constraints and transforming them to expressions in monadic second-order logic that express how to reach the auxiliary labels mentioned above.

Finally, the shrunken set of trees are glued together to restore sharing information lost in the first phase: multiple nodes that have been annotated as associated with the same pointer expression are identified, and nodes that were annotated as being shared heap nodes are made into shared nodes. The normalization results are used in a crucial way in this glueing stage, since the glueing can create many new paths within the represented store. Glueing cannot, however, create new paths through the heap in any store, since the glueing process only identifies nodes associated with pointer expressions mentioned in the formula (or in unshared paths leading to such nodes). Since normalization implies that we are only concerned with preserving the existence and nonexistence of paths that lie strictly within the heap, this is sufficient. $\square$

Figures 6 and 7 show how the proof might work for the formula Φ:

$$\Phi \stackrel{\text{def}}{=} x\langle \mathtt{car}^*\rangle y \wedge x\langle(\mathtt{cdr.cdr})^*\rangle y \wedge \neg x\langle(\mathtt{cdr.cdr.cdr})^*\rangle y \wedge y\langle \mathtt{car}^*\rangle z \wedge y\langle(\mathtt{cdr.cdr.cdr})^*\rangle z.$$

We start with a store in Figure 6 that satisfies Φ, and then prune it into a set of trees. The auxiliary labels y' and y'' keep track of the fact that these nodes in the tree must at some point be pointed to by y. In Figure 7, the trees are decreased in size, while preserving analogs of the reachability statements: e.g., the node labeled y can reach a copy of the node z with a $(\mathtt{cdr.cdr.cdr})^*$ path, and x cannot reach a copy of y with a $(\mathtt{cdr.cdr.cdr})^*$ path. In the final stage, the tree-like model is glued together to form a traditional store that satisfies Φ.

7 Related Work

Jensen et al. have also defined a decidable logic for describing properties of linked data structures [JJNS97]. It is interesting to compare the two approaches:

- The logic of Jensen et al. allows quantifications on pointer expressions, which is forbidden in L_r. Instead, L_r allows stating both sharing constraints and allocation constraints. However, both of these can be encoded using their logic:

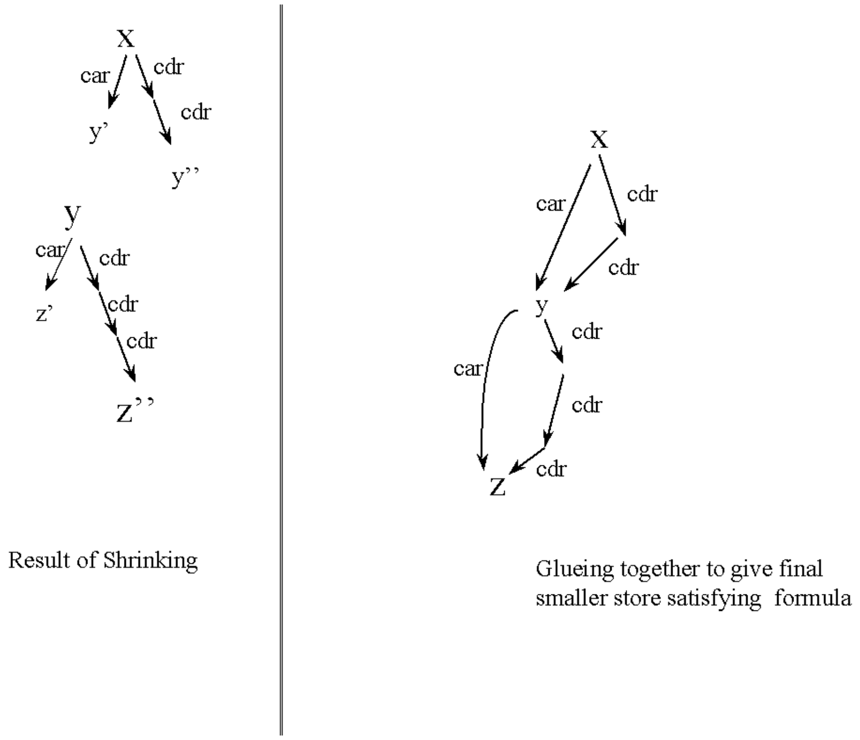

Figure 7: The Shrinking and Glueing Stages of The Proof

- Sharing constraints that can be encoded using quantifications.
- Allocation constraints can be encoded using tests for `NULL` in routing expressions.

- L_r imposes a limitation on routing expressions by forbidding testing for `NULL` and for garbage cells.
- On the other hand, L_r generalizes the logic of Jensen et al. in the following ways:
  - L_r allows multiple selectors, which enables L_r formulae to describe properties of general directed graphs as opposed to just lists.[1]
  - The reachability constraints in L_r formulae allow one to test simultaneous pointer inequities, which is crucial for capturing the strength of the variant of static shape graphs defined in [SRW98b].

In summary, the formulae of Jensen et al. are more expressive than L_r formulae, but they can only state properties of lists and trees, whereas L_r can state properties of arbitrary graph data structures.

Klarlund and Schwartzbach defined a language for defining *graph types*, which are tree data structures with non-tree links defined by auxiliary tree-path expressions [KS93]. In the application they envision, a programmer would be able to declare variables of a given graph type, and write code to mutate the "tree backbone" of these structures. After a mutation operation, the runtime system would automatically apply an update operation that they define, which updates the non-tree links. The graph-type definition language is unsuitable for describing arbitrary store graphs, and the fact that the update operations are limited does not allow the programmer to write arbitrary pieces of code. (However, the latter property is a significant advantage for the intended application — a programming language supporting controlled destructive updating of graph data structures.)

The ADDS formalism of Hendren et al. is an annotation language for expressing loop invariants and pre- and post-conditions of statements and procedures [HHN92]. From a programmer's point of view, an advantage of a logic like L_r over ADDS is that L_r is strong enough to allow stating properties of the kind that arise at intermediate points of a procedure, when a data structure is in the process of being traversed

[1] [JJNS97] sketches an extension of their technique to trees, which involves multiple selectors, but they do not handle general directed graphs.

or destructively updated. For example, ADDS cannot be used to state the loop invariant of Example 2.2 because the relationship between x and y cannot be expressed. Because it is lacking in expressive power, ADDS is mainly useful as a documentation notation for type definitions, function arguments, and function return values. Hendren et al. propose to handle this limitation of ADDS by extending it with the ability to use a certain limited class of reachability properties between variables (of the kind used in the path matrices defined in [Hen90]).

L_r goes beyond ADDS in the following ways:

- L_r permits stating properties of cyclic data structures.
- The routing expressions used in L_r formulae are general regular expressions (with pointer-variable interrogations).
- L_r is closed under both conjunction and negation. In contrast, ADDS cannot express the loop invariant in Example 2.3 because of the implication.

It should be noted that currently the notion of heap sharing in L_r is weaker than the ADDS notion of "dimension". It is easy to generalize L_r to include this concept without affecting its decidability. We did not do so in this paper because we wanted to stay with two selectors.

Finally, it should be noted that both L_r and ADDS do not allow stating connectivity properties of the form

$$\mathtt{x}\langle R_1 \rangle = \mathtt{y}\langle R_2 \rangle.$$

We believe that L_r can be generalized to handle this. (A limited form of such connectivity properties, restricted to be in the form

$$\mathtt{x}\langle (\mathtt{car}|\mathtt{cdr})^* \rangle = \mathtt{y}\langle (\mathtt{car}|\mathtt{cdr})^* \rangle,$$

was proposed in [GH95, GH98].)

L_r is incomparable to Deutsch's symbolic aliases [Deu92, Deu94]:

- Symbolic aliases allow the use of full-blown arithmetic, which cannot be used in a decidable logic.
- On the other hand, symbolic-alias expressions are not closed under negation. For instance, there is no way to express must-alias relationships using symbolic aliases. Thus, the loop invariant used in 2.3 cannot be expressed with symbolic aliases.

In [FM97], Fradet and Le Métayer use graph grammars to express interesting properties of the data structures of a C-like language. Graph grammars can be a more natural formalism than logic for describing certain topological properties of stores. However, graph grammars are not closed under intersection and negation, and problems such as the inclusion problem are not decidable. In terms of expressive power, the structure-description formalism of [FM97] is incomparable to the one proposed in the present paper.

The logic L_r is built on ideas from modal and temporal logic, and from automata-theoretic formalisms that are closely related to these (e.g., tree grammars). Negations of reachability constraints are a simple variant of the safety properties expressible in temporal logic. L_r does differ from these logics in several technical respects. L_r allows reachability statements to include regular expressions — this is a convenient language for our application. Modal logics generally lack this, although many can express these reachability constraints through other means. Most modal and temporal logics generally deal only with the reachability of certain labels in a labeled graph: they do not allow one to state, for example, that a particular node label appears only once — a property that plays a large implicit role in our logic (pointer expressions can only point to one node in a graph). Finally, these logics do not allow one to state constraints about sharing or the reachability/unreachability of shared nodes.

In fact, modal logic is well-known to possess the *tree model property*: every satisfiable sentence has a (possibly infinite) tree as a model (see, for example, the summary in [Var97]). This property fundamentally prevents modal logics as they stand from capturing interesting topological properties of graphs. The constructs given in L_r do allow one to capture topological constraints on a graph, and the framework admits extensions to a number of other topological properties as well: for example, constraints specifying the reachability of cycles can easily be accommodated in our framework.

At the level of decidability proofs, the gap between our logic and modal and temporal logics is fairly small. Our decidability proof reduces satisfiability questions in L_r to analogous questions about traditional logics on trees. In the process, it also shows that L_r has a property analogous to the tree model property: roughly, that any satisfiable sentence is satisfiable by a quotient of unions of trees. In the presence of the cyclicity constraints mentioned above, this latter property would remain true.

It should be noted that the approach given here is limited in several ways:

- The approach we have taken is to develop decidable, logic-based languages for capturing topological properties of a broad class of linked data structures. Undecidability results in predicate logic give many hard limitations on the expressiveness of such languages: For example, no such language exists that is closed under first-order quantification and boolean connectives.
- Although logic-based formalisms can be more succinct in expressing properties of linked data structures, they can also be more verbose; in particular, the output from our translation algorithms can be significantly more verbose than the input. For example, with the translation from a static shape graph SG into L_r formula $\Phi[SG]$ given in Section 4.2, the size of $\Phi[SG]$ can be exponential in $|SG|$.
- There are a few properties that cannot be expressed in L_r, including:
  - Whether a store contains a garbage cell (i.e., a cell not accessible from any variable).
  - Whether a tree is balanced (or almost balanced, such as the condition used in AVL trees).

  It may be difficult to extend L_r to handle these sorts of properties. However, such properties go well beyond the scope of current optimizing compilers and tools for aiding software understanding.

References

[AW93] U. Assmann and M. Weinhardt. Interprocedural heap analysis for parallelizing imperative programs. In W. K. Giloi, S. Jähnichen, and B. D. Shriver, editors, *Programming Models For Massively Parallel Computers*, pages 74–82, Washington, DC, September 1993. IEEE Press.

[BGG96] E. Boerger, E. Graedel, and Y. Gurevich. *The Classical Decision Problem.* Springer-Verlag, 1996.

[CWZ90] D.R. Chase, M. Wegman, and F. Zadeck. Analysis of pointers and structures. In *SIGPLAN Conference on Programming Languages Design and Implementation*, pages 296–310, New York, NY, 1990. ACM Press.

[Deu92] A. Deutsch. A storeless model for aliasing and its abstractions using finite representations of right-regular equivalence relations. In *IEEE International Conference on Computer Languages*, pages 2–13, Washington, DC, 1992. IEEE Press.

[Deu94] A. Deutsch. Interprocedural may-alias analysis for pointers: Beyond k-limiting. In *SIGPLAN Conference on Programming Languages Design and Implementation*, pages 230–241, New York, NY, 1994. ACM Press.

[FM97] Pascal Fradet and Daniel Le Metayer. Shape types. In *ACM Symposium on Principles of Programming Languages*, New York, NY, 1997. ACM Press.

[GH95] R. Ghiya and L.J. Hendren. Connection analysis: A practical interprocedural heap analysis for C. In *Proc. of the 8th Intl. Work. on Languages and Compilers for Parallel Computing*, number 1033 in Lecture Notes in Computer Science, pages 515–534, Columbus, Ohio, August 1995. Springer-Verlag.

[GH98] R. Ghiya and L. Hendren. Putting pointer analysis to work. In *ACM Symposium on Principles of Programming Languages*, New York, NY, 1998. ACM Press.

[Hen90] L. Hendren. *Parallelizing Programs with Recursive Data Structures.* PhD thesis, Cornell University, Ithaca, N.Y., Jan 1990.

[HHN92] L. Hendren, J. Hummel, and A. Nicolau. Abstractions for recursive pointer data structures: Improving the analysis and the transformation of imperative programs. In *SIGPLAN Conference on Programming Languages Design and Implementation*, pages 249–260, New York, NY, June 1992. ACM Press.

[HN90] L. Hendren and A. Nicolau. Parallelizing programs with recursive data structures. *IEEE Transactions on Parallel and Distributed Systems*, 1(1):35–47, January 1990.

[HPR89] S. Horwitz, P. Pfeiffer, and T. Reps. Dependence analysis for pointer variables. In *SIGPLAN Conference on Programming Languages Design and Implementation*, pages 28–40, New York, NY, 1989. ACM Press.

[JJNS97] J.L. Jensen, M.E. Joergensen, N.Klarlund, and M.I. Schwartzbach. Automatic verification of pointer programs using monadic second-order logic. In *SIGPLAN Conference on Programming Languages Design and Implementation*, 1997.

[JM81] N.D. Jones and S.S. Muchnick. Flow analysis and optimization of Lisp-like structures. In S.S. Muchnick and N.D. Jones, editors, *Program Flow Analysis: Theory and Applications*, chapter 4, pages 102–131. Prentice-Hall, Englewood Cliffs, NJ, 1981.

[JM82] N.D. Jones and S.S. Muchnick. A flexible approach to interprocedural data flow analysis and programs with recursive data structures. In *ACM Symposium on Principles of Programming Languages*, pages 66–74, New York, NY, 1982. ACM Press.

[KS93] N. Klarlund and M. Schwartzbach. Graph types. In *ACM Symposium on Principles of Programming Languages*, New York, NY, January 1993. ACM Press.

[LH88] J.R. Larus and P.N. Hilfinger. Detecting conflicts between structure accesses. In *SIGPLAN Conference on Programming Languages Design and Implementation*, pages 21–34, New York, NY, 1988. ACM Press.

[LM96] C.-K. Luk and T.C. Mowry. Compiler-based prefetching for recursive data structures. In *Proceedings of the Seventh International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 222–233, October 1996.

[PCK93] J. Plevyak, A.A. Chien, and V. Karamcheti. Analysis of dynamic structures for efficient parallel execution. In U. Banerjee, D. Gelernter, A. Nicolau, and D. Padua, editors, *Languages and Compilers for Parallel Computing*, volume 768 of *Lecture Notes in Computer Science*, pages 37–57, Portland, OR, August 1993. Springer-Verlag.

[SRW98a] M. Sagiv, T. Reps, and R. Wilhelm. Parametric shape analysis via 3-valued logic. Technical Report TR-1383, Computer Sciences Department, University of Wisconsin, Madison, WI, July 1998. Available at "http://www.cs.wisc.edu/wpis/papers/parametric.ps".

[SRW98b] M. Sagiv, T. Reps, and R. Wilhelm. Solving shape-analysis problems in languages with destructive updating. *ACM Transactions on Programming Languages and Systems*, 20(1):1–50, January 1998.

[Str92] J. Stransky. A lattice for abstract interpretation of dynamic (Lisp-like) structures. *Information and Computation*, 101(1):70–102, November 1992.

[Tar81a] R.E. Tarjan. Fast algorithms for solving path problems. *J. ACM*, 28(3):594–614, 1981.

[Tar81b] R.E. Tarjan. A unified approach to path problems. *J. ACM*, 28(3):577–593, 1981.

[TW68] J.W. Thatcher and J.B. Wright. Generalized finite automata with an application to a decision problem of second order logic. *Math. Syst. Theory*, 2:57–82, 1968.

[Var97] M. Vardi. Why is modal logic so robustly decidable? In N. Immerman and P. Kolaitis, editors, *Descriptive Complexity and Finite Models*. AMS, 1997.

[Wan94] E. Y.-B. Wang. *Analysis of Recursive Types in an Imperative Language.* PhD thesis, University of California, Berkeley, CA, 1994.